The promise of a DNA taxonomy

THE ROYAL SOCIETY

Published online 22 March 2004

The promise of a DNA taxonomy

Mark L. Blaxter

Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK (*mark.blaxter@ed.ac.uk*)

Not only is the number of described species a very small proportion of the estimated extant number of taxa, but it also appears that all concepts of the extent and boundaries of 'species' fail in many cases. Using conserved molecular sequences it is possible to define and diagnose molecular operational taxonomic units (MOTU) that have a similar extent to traditional 'species'. Use of a MOTU system not only allows the rapid and effective identification of most taxa, including those not encountered before, but also allows investigation of the evolution of patterns of diversity. A MOTU approach is not without problems, particularly in the area of deciding what level of molecular difference defines a biologically relevant taxon, but has many benefits. Molecular data are extremely well suited to re-analysis and meta-analysis, and data from multiple independent studies can be readily collated and investigated by using new parameters and assumptions. Previous molecular taxonomic efforts have focused narrowly. Advances in high-throughput sequencing methodologies, however, place the idea of a universal, multi-locus molecular barcoding system in the realm of the possible.

Keywords: DNA barcodes; molecular taxonomy; diversity; operational taxonomic units

> ... endless forms most beautiful and most wonderful have been, and are being, evolved.
>
> (Darwin 1859)

1. ENDLESS FORMS: THE TAXONOMY DEFICIT

It is widely recognized that not only is the biotic diversity on planet Earth currently undergoing a mass extinction, but that the true extent of the extinction is unknown (May 1988). Although prominent marker taxa such as endangered vertebrates are well studied, most taxa remain to be discovered (Blaxter 2003). It is estimated that there are *ca*. 1.5 million described taxa (at the species level; de Meeus & Renaud 2002). However, in most major groups (phyla or divisions) the number of described taxa is a small proportion of the estimated diversity. Thus the true species-level diversity of nematodes, with *ca*. 26 000 described species, is estimated to be in the low millions (Lambshead (1993); but see also an alternative view expressed in Lambshead & Boucher (2003)), whereas only *ca*. 25% of arthropods have been described, despite a 'known' species count of over one million.

By examining the current rates of description of new taxa, the gaps in our knowledge can be split into two types. At the species and genus level, we are ignorant of most diversity in most taxa. Above the generic level, discovery of new families, orders and phyla is increasingly rare. Novel insect orders are a matter for wide discussion (Klass *et al.* 2002), and novel animal phyla have been identified only in the past 100 years by the inestimable R. M. Kristensen, who has described two (Kristensen 1983; Funch & Kristensen 1995). New higher taxa are usually small and rare, or at least unculturable. Thus, many new prokaryotic divisions (the systematic bacteriology analogue of phyla) have been described in recent years in culture-independent studies (Pace 1997). For some taxa, a relatively robust estimate of the depth of our ignorance can be made: the vertebrates are unlikely to have more than 10% undescribed taxa. For others, however, we do not even know the extent of our ignorance, and estimates of 'true' species-level diversity are mere guesses with wide confidence limits. Thus, for nematodes estimates range from 40 000 to 100 million (Blaxter *et al.* 1998).

A confounding problem in counting the diversity of taxa is that the basic grouping, the species, does not have a single definition. There are many different philosophical concepts of 'the species' that often conflict, and all fail in some cases (Adams 1998, 2001). Thus, biologically based (interbreeding) concepts fail for all asexual lineages, or for specimens pickled in a jar, and morphological ones fail for reproductively isolated but recently separated sibling taxa. The issue has been concisely summarized by Byron Adams as 'the species delimitation uncertainty principle' (Adams 2001): the more closely and precisely a species concept is defined, the less possible it becomes to diagnose whether a particular individual or specimen is a member of that species. Expanding ideas of 'true diversity' have historically been matched by a reducing pool of systematists able to diagnose that diversity, and by the realization that the taxonomic effort to classify the hyper-abundant small taxa is far in excess of any currently available workforce (Lawton *et al.* 1998). This has led to recurring cries of crises in taxonomy, and of the need for a rejuvenation of the field (Godfray 2002*a*,*b*). Below, I explore the possibilities and promise of the introduction of a genetic marker system, a DNA barcode, into taxonomic

One contribution of 19 to a Theme Issue 'Taxonomy for the twenty-first century'.

initiatives. An impassioned and well-argued 'plea for DNA taxonomy', which should be consulted for alternative views on this topic, has been made by Tautz *et al.* (2002, 2003).

2. THE PROMISE OF DNA BARCODES

The product barcode has become a universal feature of modern life. A barcode is a machine-readable digital tag, usually a series of stripes, which encodes information about the item to which it is attached. There are several alternative barcode systems that all have the same features: they identify items to a useful level of uniqueness. In a hospital this is to an individual human, whereas in a supermarket it is to one instance of a multimillion-member class. The barcode can also include some systematic or 'taxonomic' information, yielding data not only on type but also on attributes such as origin, major classification and date. Thus an apple may, in addition to a tag indicating what sort of apple it is, have a major classification 'fresh produce', a 'best by' date and a supplier code. A similar universal system is used in publishing, where the ISBN not only uniquely identifies the book, but also the publisher and the edition.

The genomes of living organisms are analogous to barcodes. Despite functional constraints, there is ample information space in a genome for complex records of individual identity and group membership. In the human population, it is estimated that each unrelated pair of individuals will differ at *ca*. 0.1% of their DNA bases (across the 3 gigabase genome this amounts to 3×10^6 differences). However, these within-taxon differences are not randomly distributed: they cluster in the less important parts of the genome—in the third, wobble bases of codons, and in intronic and intergenic DNA. There are therefore significant stretches of the genome that are maintained by selection to be (near-)identical between members of a taxon, but which can vary between taxa. It is these segments that are most useful for identification and taxonomy. As sequences evolve, they maintain records of their deep pasts as well as markers of their more recent history. A DNA barcode, derived from the sequence of a part of the genome of the organism, could in theory carry both specific and systematic data. This evolutionary aspect qualitatively differentiates DNA barcodes from product barcodes: the data they encode about relationships are retained through evolution in a stochastic fashion, rather than being hard-coded for utility by a rational agent.

There exists an alphabet soup of methods for generating a molecular fingerprint from an organism (restriction fragment length polymorphisms, amplified fragment length polymorphisms, denaturing gradient gel electrophoresis). These are generally based on assessment of length differences between fragments tagged by the presence of a short sequence (such as a restriction enzyme site or the presence of a repetitive sequence element). Although these fingerprints do yield barcode-like data, they are less than optimal for a molecular taxonomy because of problems with high within-taxon variability and lack of confident assignment of orthology between markers. DNA sequences can overcome these hurdles. A candidate DNA barcode sequence target must be known to be orthologous between specimens, as paralogues will define gene not organismal relatedness, and must encompass sufficient variability to allow discrimination between taxa useful to the research programme. The Darwinian relatedness of organisms means that there are many candidate genomic sections available. Many genes perform core functions in life processes yet vary between individuals. Some regions of the genome are sufficiently conserved to allow the use of 'universal' oligonucleotide primer sets for PCR amplification but also contain informative sequence difference. These target genes are extremely unlikely to be functionally involved in the process of speciation, and thus genetic variation between individual organisms is not directly attributable to taxon status. However, as lineages diverge in phylogenetic time, random fixation of mutations (both those present in the shared ancestor of the populations and those arising in the time since separation) will result in the accumulation of fixed differences. Thus a universal DNA barcode marker will not, in most cases, be able to distinguish very recently separated taxa (Verheyen *et al.* 2003).

To be clear that what is being estimated for a specimen is not necessarily its membership of a 'species', however defined, we call the taxa yielded by grouping of specimens through a set of markers OTU. We have coined the term MOTU (Floyd *et al.* 2002); MOTU have also been called 'phylotypes' and 'genospecies'. MOTU can be simply defined by sequence identity: if two specimens yield sequences that are identical within some defined cut-off, they are assigned to the same MOTU. However, it is important to note that MOTU membership of a specimen need not correspond to its membership of any other OTU, measured by other models (biological or morphological). This problem is not unique to MOTU, as all methods must use some diagnostic heuristic, which may result in incongruence in OTU circumscription. However, for MOTU this problem is neatly definable in terms of the level of sequence identity used in their definition: if a researcher thinks that the rules used were too lax or too strict, they can simply acquire and reanalyse the data. Given that it is clear from many gene sequences that different higher taxonomic groups can differ markedly in their background and adaptive substitution rates, and that different sized populations might be expected to harbour different levels of within-taxon variation (also dependent on the populations' evolutionary history), it may be necessary to define different heuristics for MOTU designation depending on the higher taxon studied.

3. THE SPECIAL PROPERTIES OF DNA BARCODES

With a sequence-based molecular taxonomy, a single technique is applicable to all taxa: extract DNA, PCR and sequence. A standard protocol for DNA barcode determination is simple to devise and promulgate, and can be applied on a high-throughput scale (see figure 1 for a summary of a DNA barcoding system). It is not necessary to have a specific training in the nuances of the taxonomy of the group of interest. Complete data can be obtained from single specimens irrespective of sexual morph or life stage, often without compromising parallel or subsequent morphological identification. Morphologically indistinguishable taxa can be diagnosed without the need for live material,

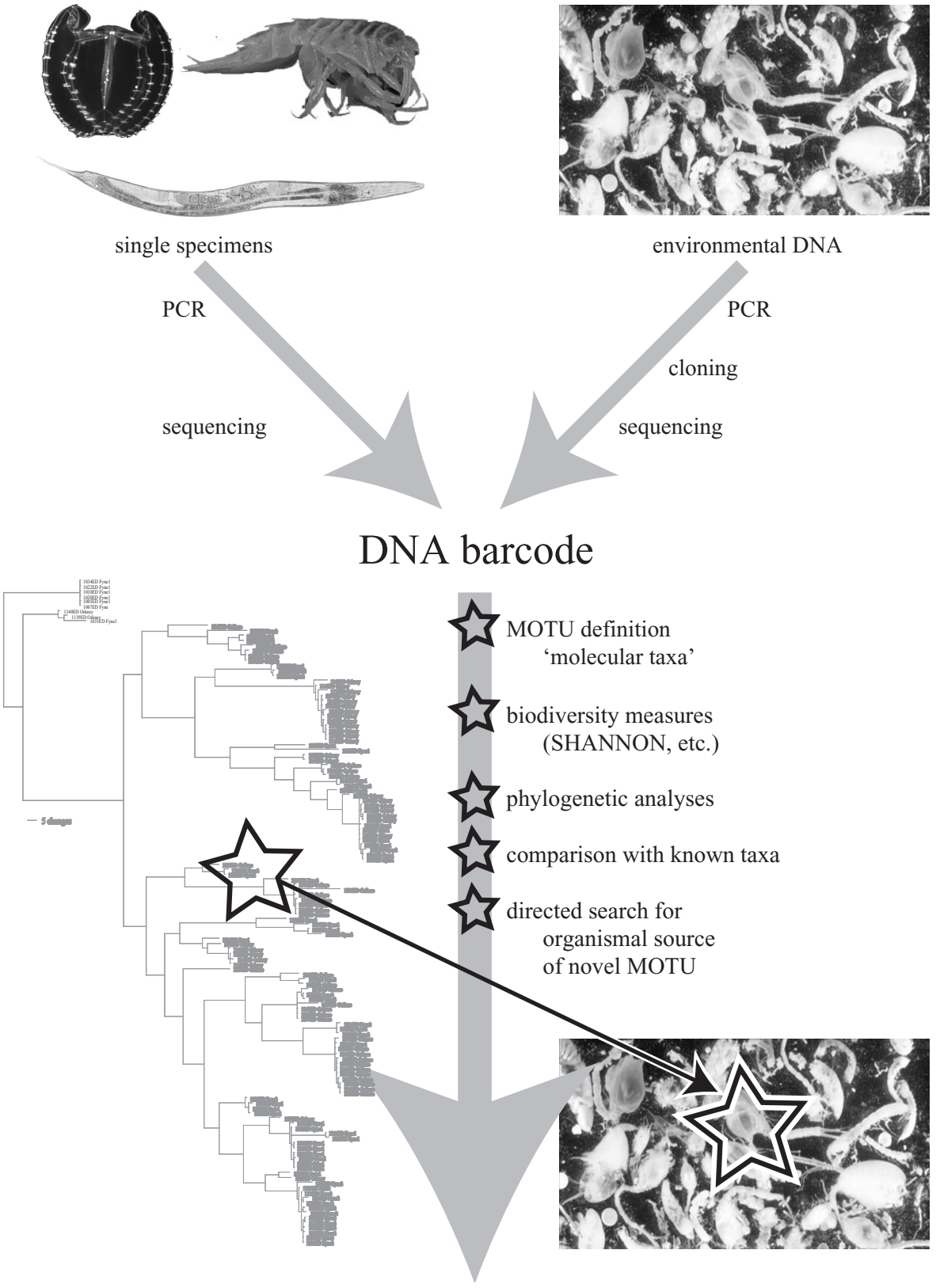

Figure 1. An overview of the process of DNA barcoding for taxon identification. This cartoon illustrates an overall strategy that could be followed in a molecular barcoding and taxonomy programme. By using PCR gene fragments, the DNA barcode targets can be isolated either from single specimens with digital images or preserved parts as vouchers, or from DNA extracted from a community of organisms ('environmental DNA'). The sequences generated constitute the barcodes for that specimen, or for the unidentified organism. The barcode sequences can be compared with each other to define membership of MOTU, and thus circumscribe taxa. The sequences can also be compared with orthologous barcode sequences obtained from specimens previously identified to taxon by other means, and thus some correspondence between MOTU and other taxonomic systems achieved. In the case of barcodes derived from 'environmental DNA', the discovery of novel sequence types can lead to a directed prospecting for the originating organism (see also Oren 2004).

particular morphs or population measures (see also Gaston & O'Neill 2004). Finally, previously unencountered taxa are as easy to analyse as dominant taxa, and the perils of accidental synonymy or elision are much reduced.

Most importantly, the raw data of the taxonomic diagnosis process, the sequences, can be used to place MOTU in the context of the rest of biology. We stand on the shoulders of giants, workers who have diligently described and summarized the biology and ecology of thousands of species. These data are crucial to deriving biology from sequence. Barcode sequences can be generated from type specimens (holotype, paratype or neotype) under strict conditions of traceability and verification. These stepping-stone sequences provide a route between the MOTU and traditional systems. A specimen barcode can be compared with sequences derived from other molecular taxonomy initiatives. If a close match is found to a named taxon, recourse can be made to traditional monographs and keys to understand the biological properties of the identified MOTU and their close relatives (Floyd *et al.* 2002) and molecular phylogenetic analyses can be used to generate testable hypotheses of MOTU interrelatedness.

The sequences used thus far for molecular barcoding are the nuclear small subunit ribosomal RNA gene (SSU, also known as 16S in prokaryotes, and 18S in most eukaryotes), the nuclear large-subunit ribosomal RNA gene (LSU, also known as 23S and 28S; the highly variable expansion loops that are flanked by conserved stem sequences are particularly useful), the highly variable internal-transcribed spacer section of the ribosomal RNA cistron (ITS, separated by the 5S ribosomal RNA gene into ITS1 and ITS2 regions), the mitochondrial cytochrome *c* oxidase 1 (CO1 or COX1) gene and the chloroplast ribulose bisphosphate carboxylase large subunit (*rbcL*) gene.

The relative utility of these can be rated by: (i) their ease of isolation from a sample; (ii) the likelihood of within-individual variation; (iii) the ease of alignment and analysis; (iv) the number of sequences already known from identified specimens; and (v) the potential universality of a barcode based thereon. For all these targets, PCR facilitates amplification from even single cells, and their multicopy nature is also advantageous. Sequencing is essentially equally easy for all DNA fragments barring extreme base composition biases, polynucleotide runs and stable secondary structures. However, the ITS region often varies by insertions or deletions within an individual, making sequencing very difficult (as two independent sequence types are being analysed simultaneously) (Elbadri *et al.* 2002); ITS sequences are also very difficult to align as they tend to evolve by insertion and deletion rather than substitution, making the secondary steps of phylogenetic reconstruction problematic. SSU, LSU, COX1 and *rbcL* are each relatively simple to align and analyse, though exceptions do occur. For example, the SSU of the rhabditid nematode *Pelodera strongyloides* has regions that allow facile alignment with related taxa such as *Caenorhabditis elegans* interspersed with segments of apparently randomized bases (Fitch *et al.* 1995). The protein-coding genes COX1 and *rbcL* have the pleasing property of being reliably partitioned into codons with first, second and third base positions. Within these, fourfold redundant sites in third base positions will be essentially neutrally evolving, and thus will have a higher rate than twofold redundant sites, first bases or second bases: these partitions allow analysis at many different levels of divergence.

In terms of existing databases for these sequences, the ITS and SSU are best represented with over 30 000 sequences each; for COX1 and *rbcL* there are *ca.* 17 000 sequences each. SSU and LSU are truly universal genes, found in every organism, and there are excellent and well-tested primer sets that work on most taxa. The ITS region is peculiar to Metazoa, but can be amplified by using primers derived from flanking SSU and LSU genes, and is thus universally amplifiable. As COX1 is derived from the proteobacterial symbiont that gave rise to the mitochondrion, it also has universal presence, but primer sets for amplifying COX1 fragments tend to be less universally applicable. *rbcL*, as a chloroplast gene, is only useful for Viridiplantae and related taxa. For these organellar genes there are problems associated with horizontal acquisition of organelles through hybridization events (Bergthorsson *et al.* 2003). From the above, I would suggest that any barcoding system should aim to acquire data for at least a nuclear and an organellar gene from single specimens. For specimen-independent, 'environmental DNA'-based surveys, any target may do, but the universality of SSU and LSU primer sets recommends them.

The sequence data from a barcoding experiment is not the only important output: the base string is derived from a sequencing experiment performed on a DNA extract from a specimen that may have remaining morphology or have associated digital vouchers (De Ley & Bert 2001). Thus, it is important, when considering the storage of barcode data, to also consider long-term storage of the specimens and images, and the DNA extracts (for example desiccated and frozen on paper discs). These may allow later addition of more data from additional genes. The sequence experiment data, usually a fluorescent sequencer chromatogram, can also be archived electronically. The chromatogram (or base quality scores derived from it; Ewing & Green 1998; Ewing *et al.* 1998) can indicate the level of support each base called has, and thus assist in discriminating between error and biology. In my opinion it is important to archive all sequences, and not just one representative or consensus from each MOTU found in a survey. This will allow later meta-analyses across datasets using standardized parameters. The Internet is the perfect medium for this (Bisby *et al.* 2002; Godfray 2002*a*; Lee 2002; Oren & Stackebrandt 2002), particularly perhaps through a carefully designed and implemented taxonomic annexe to the central sequence databases of the European Bioinformatics Institute, National Center for Biotechnology Information and DNA Database of Japan (though there are problems with this approach; Tautz *et al.* 2003). Examples of how such a universal database might function are available in the form of the European and US ribosomal RNA database projects (Wuyts *et al.* 2001, 2002; Cole *et al.* 2003), which each aim to collect, align and make available SSU and LSU sequences from all taxa. Importantly, the US 'RDP-II' (Cole *et al.* 2003) has a focus on the analysis of environmentally sampled, culture-independent datasets (see http://rdp.cme.msu.edu/html/). A project to promote DNA taxonomy has also been initiated in Munich (see http://www.zsm.mwn.de/

DNATAX/) and a DNA barcoding Web site focusing on the COX1 gene is under development (see http://www.barcodinglife.com/) (Hebert *et al.* 2003*a*,*b*).

4. PROBLEMS WITH BARCODES: WHEN DOES A SEQUENCE MEAN A TAXON?

> From these remarks it will be seen that I look at the term species, as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other, and that it does not essentially differ from the term variety, which is given to less distinct and more fluctuating forms. The term variety, again, in comparison with mere individual differences, is also applied arbitrarily, and for mere convenience sake.
>
> (Darwin 1859)

Differences in sequences between specimens can arise in several ways. They could be due to methodological errors, be part of the natural, within-OTU variation or be related to a distinction between taxa. It is thus necessary (as with other methods, biological or morphological) to use heuristics for MOTU distinction based on known error rates in measurement, and perceived levels of difference that distinguish 'useful' OTU. For MOTU, these measures can be made explicit. For example, from known, accepted taxa within a particular group, the level of natural variation present within a taxon (say a breeding population of organisms) can be measured and compared with the difference observed between populations, or between taxa.

Both nuclear and organellar genomes can be 'heterozygous', and the multiple copies of target genes in the genome can differ in sequence. For example *Plasmodium falciparum*, the malaria parasite, has multiple different ribosomal RNA cistrons that are differentially expressed in the complex life cycle (Mercereau-Puijalon *et al.* 2002). Organellar genomes can be heteroplasmic. A useful barcode marker will have very low within-population and between-population diversity, and measurable between-taxon diversity.

Multiple resequencing of a single specimen can be carried out to assess both the PCR-generated and sequencing technology-generated error rate. It is likely that these methodological errors will not always be sequence-independent, as the biochemistry of thermostable polymerases is sequence environment sensitive. A comparison between the between-taxon difference rate and the within-taxon variation and error rates will define the accuracy and specificity of MOTU definition. Sequencing directly from PCR products rather than from clones of PCR amplicons eliminates most PCR-introduced error, and in particular eliminates the problem of chimaeric clones resulting from between-amplicon priming. Current sequencing methodologies yield sequences that are usually highly accurate over *ca*. 500 bases (less than 1 error in 10 000), and thus the experimental error rate is usually much less than natural variation.

A major issue still to be resolved is how to derive MOTU from barcode sequences. As not all variation between sequences will be biologically relevant (i.e. some will be experimental error, and some will be within-taxon variation), simply classifying every unique sequence type as a biological OTU is not biologically realistic. Similarly, using a simple percentage difference from a series of reference sequences, as is the norm in prokaryote studies (see below) is also not optimal. In our analyses of nematode DNA barcodes (see below) we (Floyd *et al.* 2002; Blaxter *et al.* 2003) have developed a series of informatic techniques that address issues of sequence quality, sequence length and between-sequence difference to generate MOTU (clusters of sequences) at any level of base pair difference desired. We have noted that whatever level of difference is selected as the discriminant, re-running a clustering process will not necessarily yield the same set of MOTU. The same problem also arises in many sequence-clustering issues in genomic biology (Parkinson *et al.* 2002). Thus, with a taxon delimitation threshold of two base differences, a pair of sequences that each differ by two bases from a third sequence, but by four bases from each other, could be grouped into one or two taxa depending on the order in which they are analysed. In recognition of this we would propose that any clustering into MOTU is based on repeated cycles of random addition of sequences, and that the resulting MOTU are recognized as instances of a stochastic process. Thus, the importance of archiving the raw sequence data as well as the derived OTU arises; with the raw data, a re-analysis is possible. Other methods of analysis, such as using multi-dimensional scaling to cluster sequences (Hebert *et al.* 2003b), are also viable.

The resolution of these issues will only come from extensive testing of DNA barcode MOTU versus other sorts of biological taxonomic unit identification, using real, wild populations as substrates. Are MOTU discovering taxa that traditional biology would recognize? Are traditional taxa recovered as MOTU? Where there are discrepancies, which method better reflects the underlying biology?

5. DNA BARCODING IN PRACTICE

> The primary aims of taxonomy are to name, circumscribe, describe and classify species. The first goal is convention but the remainder are science.
>
> (Seberg *et al.* 2003, p. 63)

From the above, I suggest that a DNA barcode system is likely to be able to achieve the three scientific goals of taxonomy defined by Seberg *et al.* (2003) and thus support a broad spectrum of taxonomic and systematic studies. But do DNA barcodes work in practice? The answer is a resounding yes, but some work remains before a barcode system is likely to become truly universal.

Ten years ago, the first surveys of bacterial diversity using culture-independent methods shocked the world of prokaryotic systematics. By amplifying and sequencing 16S SSU genes amplified from 'environmental DNA' extracted from sieved-out microbes (Giovannoni *et al.* 1990; Fuhrman *et al.* 1992), a new view of the diversity of the microbial world emerged (Woese 1996; Pace 1997; Hugenholtz *et al.* 1998). In all environments, the numbers of identifiably different SSU sequences was 20–100-fold greater than those measured in culture-based studies. These new sequences were sometimes sufficiently close to known cultured taxa, but in many cases they suggested deeply divergent lineages with no known cultured

representatives. Many of these lineages have subsequently been shown to be widespread, in that members have been sequenced from many different environments, and sometimes dominant, either numerically or ecologically. The rate of discovery of new 'domains' has slowed, but even apparently impoverished habitats still yield new sequences. Terrestrial soils from temperate (Furlong *et al.* 2002) or tundra (Zhou *et al.* 1997), lakes (Casamayor *et al.* 2002), and even the flora of the pig gut (Leser *et al.* 2002) yield a similar story: unexpected, deep diversity. However, it is noteworthy that some globally distributed ecosystems have an overall low bacterial diversity (Hentschel *et al.* 2002).

The explosion of sequence data in bacterial typing has led to the establishment in microbial systematics of a series of conventions for analysis and naming originating from initial efforts to compare bacterial genomes by hybridization (Wayne *et al.* 1987; Stackebrandt & Goebel 1994). This initial convention has been translated into a series of heuristics for defining 'genospecies' on the basis of proportional sequence identity (Cohan 2002). The underlying concept is that, because bacteria have a clonal population structure that allows rapid fixation of sequence change, it is possible for lineages to differ in derived substitutions and yet be part of the same genospecies. However, the heuristics used by different workers are not the same: some use a 'strict' 99% identity (over *ca*. 500 bases), whereas others use a more relaxed 97% or even 95%. This difference in taxon discrimination makes comparison of different studies difficult at best. Bacteria also embody one of the more cogent arguments against a taxonomy based on 'an infinitesimally tiny fraction of an organism's genome' (Lipscomb *et al.* 2003): horizontal gene transfer (Smith *et al.* 1992; Doolittle 1999; Nesbo *et al.* 2001). The transfer of genes between two distinct taxa has two effects. The horizontally transferred genes will have a different phylogenetic history from the remainder of the genome, and thus any taxonomy or systematics based on them will be incongruent. However, even in the presence of rampant horizontal acquisition, a core genome of coevolving genes can be recognized (Daubin *et al.* 2003), and for a universal system a member of this core genome is recommended. Secondly, transferred genes are often physiologically and ecologically dominant (Van Tienderen *et al*. 2002). Variation in the resident ribosomal RNA genes will have little direct effect on the functional abilities of bacterial taxa, whereas acquired nitrogen fixation, photosynthetic or xenobiotic metabolism genes will. It may therefore be more cogent to a particular research programme to determine the molecular diversity of ecologically important genes rather than the organisms that carry them (Karp *et al*. 1997; Van Tienderen *et al.* 2002).

The success of the SSU-based bacterial diversity discovery programme has inspired workers focused on eukaryotic groups to test similar methods. From the first fruits of these studies it is evident that eukaryotic microbial diversity has likewise been underestimated (Moreira & Lopez-Garcia 2002). These studies have in the main used the 18S SSU. SSU gene libraries derived from planktonic organisms of deep sea (Diez *et al.* 2001; Lopez-Garcia *et al.* 2001; Massana *et al.* 2002), open ocean (Moon-van der Staay *et al.* 2001), deep-sea vent communities (Edgcomb *et al.* 2002) and an acidified iron-rich spring-fed river (Amaral Zettler *et al.* 2002) have been sampled by sequencing. In each case novel sequence-defined taxa have been discovered, including those that suggest new major clades. Some of these new eukaryotes are so small that they have previously been included with the prokaryotes (Diez *et al.* 2001; Massana *et al.* 2002). From these and other more directed studies, the Eukaryota can now be divided into eight domains, only two of which are familiar (fungi plus animals, green photosynthesizers including plants).

Fungi, particularly the fungi of soils and the rhizosphere, are problematic from a morphological taxonomic standpoint, as taxon diagnosis is often made by examination of the spores whereas the bulk of fungal material is hyphal. SSU-based analyses of rhizosphere fungi suggest both a greater overall diversity of taxa involved than is evident from spore analysis, and a dynamic association between plant roots and their symbionts (Vandenkoornhuyse *et al.* 2002); it is likely that surveys focused on fungi will reveal a similar diversity in other habitats. In the Viridiplantae, molecular taxon markers have been used for some time, based mainly on the chloroplast-encoded *rbcL* gene. In addition to identifying taxa (see, for example, McDaniel & Shaw 2003) *rbcL* sequences can also used to untangle species' hybridization history (Nishimoto *et al.* 2003) and to probe the deep phylogeny of the plants (Savolainen *et al.* 2000).

Application of molecular barcoding concepts to animals has been slowest off the ground, perhaps because of the obsession of most zoologists with larger taxa, and the abundance of morphology in the hyper-speciose arthropods. However, even in big animals, molecular tags are being used to define new taxa from within the confines of old, most spectacularly in the case of the 'new' species *Loxodonta cyclotis*, the African forest elephant (Grubb *et al.* 2000; Roca *et al.* 2001; Eggert *et al.* 2002). For meiofaunal organisms, the nematodes appeared to us to be a good test case and the nuclear SSU gene an ideal target (Blaxter *et al.* 1998). Nematodes are hyper-abundant (14 billion per hectare of a poor Scottish upland soil) and low on the list of animals with an exciting morphology. Within the community of nematode morphologists there is widespread recognition that many characters are homoplasious between taxa (De Ley 1999), and that a taxonomist may diagnose sexual species by only the merest asymmetry in cell position (Felix *et al.* 1996). Preliminary surveys of SSU sequence suggested that divergence between congeneric species pairs was likely to be sufficient to diagnose taxa at an appropriate species or genus level (Blaxter *et al.* 1998). A small grassland field in southern Scotland, the focus of the Natural Environment Research Council Soil Biodiversity and Ecosystem Function programme, was chosen, and nematodes sampled individually by direct sequencing of PCR products. Taxa were defined by more than 99.6% identity in sequence by using a custom sequence-clustering algorithm adapted from a genomics application (Parkinson *et al.* 2002). Culturable diversity (seven MOTU in 170 cultures derived from 1200 individual females tested; Floyd *et al.* 2002), was far below culture-independent diversity (between 135 and 150 MOTU from 2000 specimens of both sexes and all stages; Floyd *et al*. 2004). Given that there are approximately 200 known soil nematode species in the UK, this 1 ha hillside field appears to be hyper-diverse, but it is more likely that

the MOTU method is identifying taxa cryptic to morphological analysis. Importantly, we were able to test, for cultured isolates, the match between MOTU, morphological taxa and taxa as defined by reproductive compatibility: MOTU and reproductive taxa were fully congruent, whereas morphology was at best indiscriminate and at worst in conflict even between specimens from the same isofemale line (Eyualem & Blaxter 2003). For another example of a nematode barcoding project, see figure 2.

More recently, Paul Hebert and colleagues (Hebert *et al.* 2003*a*,*b*) have proposed the use of the COX1 gene as a metazoan barcode target. They have shown that COX1 sequences can correctly diagnose moth specimens to morphologically defined species, and in general, assign an unknown specimen by sequence to higher taxonomic ranks (Hebert *et al.* 2003*b*). Comparison of COX1 sequences between congeneric species pairs shows that in all but a few cases (notably the Cnidaria) the sequence divergence is greater than 2% (Hebert *et al.* 2003*a*). This bodes well for the use of COX1 as a barcode, but there may be limitations in its use for environmental DNA surveys as the 'universal' primer binding sites are missing from several major taxa (such as orders of tardigrades and nematodes).

Figure 2. (*Overleaf.*) DNA barcoding in practice: a survey of littoral nematodes. As an example of a molecular barcoding project, a set of unpublished data is presented from R. Roche and M. Blaxter, surveying the nematode communities of three Scottish beaches (a sheltered muddy beach at Loch Fyne, an exposed sandy beach in the Orkneys and an estuarine sandy beach at Gullane, shown in (*c*)). Between 30 and 35 single nematodes were isolated from each sample using standard methods. Loch Fyne was sampled twice. Nematodes were identified to order (and genus where possible) before processing for sequencing of the 5′-end of the SSU gene. The 134 resulting sequences were clustered into MOTU by using the procedure of Floyd *et al.* (2002). Fifty-one MOTU were defined containing 1–13 individuals. An illustrative phylogram is shown in (*a*), generated by using the neighbour joining algorithm using absolute differences: sequences belonging to the same MOTU are boxed. The MOTU could be ascribed to known taxa in only a few cases, owing to the paucity of sequences available from marine nematodes, but, for example, the sequences in MOTU_047 are identical to one determined from *Ascolaimus elongatus* whereas two of the specimens whose sequences were clustered in MOTU_001 were identified as *Monoposthia costata*, suggesting that all those specimens were *M. costata* or a very closely related taxon. The distribution of MOTU between sites is shown in a Venn diagram in (*b*). Each sample of *ca.* 35 nematodes yielded 12–15 MOTU. Within Loch Fyne, the two independent samples shared only four MOTU, and only two MOTU were common to Orkney and Loch Fyne, and Gullane and Loch Fyne. This low overlap suggests high within- and between-site diversity.

6. BEYOND BARCODES: SYSTEMATICS FROM DNA

Barcode DNA sequences are chosen for both conservation, permitting facile alignment between instances, and variability, allowing us to diagnose taxa. These features also make them suited to model-driven phylogenetic analysis. Although it is essential to keep in mind that what are being constructed are gene trees, these trees, we hope, bear a significant relationship to the phylogeny of the taxa under consideration. There are many reasons, both biological and methodological, why molecular phylogenetic analyses will robustly find a tree that is 'wrong' by other criteria, and DNA barcode target sequences are no less prone to these artefacts than any other. However, all the barcode targets have a long and fruitful history of use in molecular phylogeny, and many tools and algorithms are available to assess deviance from normal behaviour, and in many cases correct for it (Felsenstein 1978; Swofford *et al.* 1996; Huelsenbeck 1997).

Barcode sequences are, in general, short (*ca.* 500 bases, the length of a single sequencing run) and this fundamentally limits their utility in resolving deep branches (between orders or phyla) in phylogenies. However, they are perfect for ordering the terminal and sub-terminal nodes on trees. Despite fears to the contrary (Seberg *et al.* 2003), the conservation of the target sequences means that in most cases, alignment is unproblematic. Difficulties can arise when a sequence from a completely novel group is discovered, but this is but a spur to closer analysis rather than a fatal flaw.

Adding DNA barcodes to newly described taxa is a relatively simple task. Excitingly, it may be possible in some circumstances to extend the reach of molecular taxonomy into the past. DNA is a relatively stable molecule, and can be isolated from museum specimens stored dry or in alcohol. Isolation of 'vintage' DNA from formalin-fixed specimens is more difficult, and more destructive of the specimen (Herniou *et al.* 1998). For larger specimens, such as mammals and birds, collections made in the past few hundred years can be surveyed by sampling a few hairs, feather bases or a few milligrams of tissue. These small subsamples can yield a canonical sequence for a taxon holotype or paratype, and help address the status of extinct taxa or populations and their relationships to modern ones. For smaller, alcohol-preserved specimens, for example flatworms (Herniou *et al.* 1998), sampling may be more destructive but still retain specimen morphology. This is particularly true of arthropod taxa, where it is the sclerotized or calcified exoskeletal elements that yield morphological data and the softer tissues within that yield DNA sequences (Dabert *et al.* 2001).

A major unanswered issue with old, vintage and ancient DNA is that of age-induced changes that result in base substitution, either biased substitution (where one base is preferentially 'mutated' to another) or accelerated substitution (where the process is unbiased, but results in overall increased observed substitution rates; Herniou *et al.* 1998; Hofreiter *et al.* 2001*a*,*b*). These effects would tend to produce overestimates of taxon number in a molecular taxonomic framework but should be minimized by repeated re-PCR and re-sequencing to identify error positions.

7. THE FUTURE

Therefore, in conclusion, the idea of molecular barcoding for taxonomic purposes is already a reality (Blaxter 2003; Blaxter & Floyd 2003). Descriptions of new 'species' are being published with a DNA sequence attached

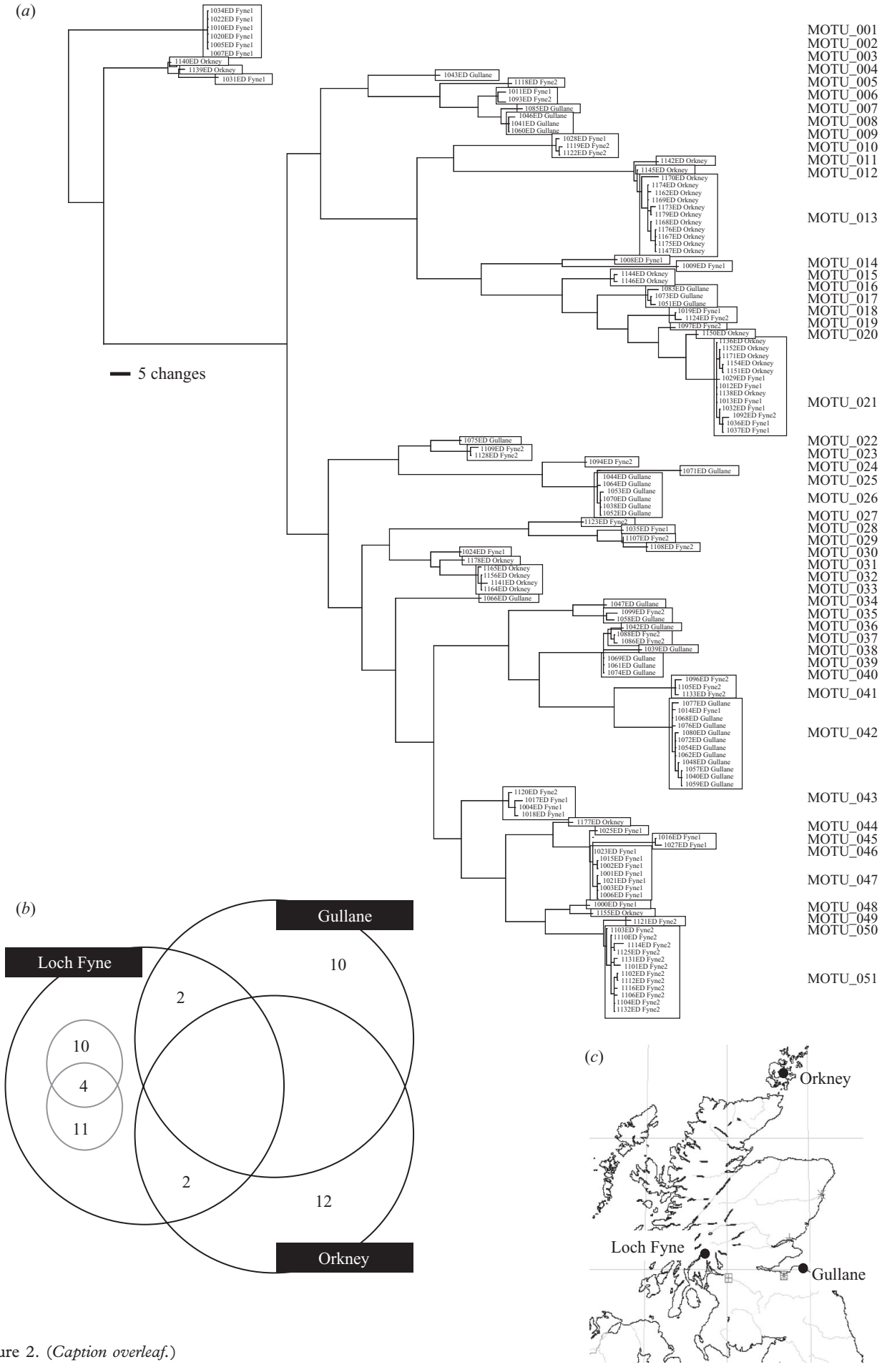

Figure 2. (*Caption overleaf.*)

to the primary nomenclatorial act (Sommer *et al.* 1996) and this should be actively encouraged. Taxonomists preparing new taxa for publication should be welcomed by DNA-savvy biodiversity laboratories that should be able to provide expertise at minimal cost. Methods for rapid sequence acquisition for minimal cost are already in existence at genome sequencing centres, and are easily adapted for taxonomic sampling. There needs to be a core plurality in the programme, with more than one sequence per specimen being perhaps a minimal aim, and a concerted attempt to link MOTU defined with one barcode target with those from other targets (for example directed sequencing of COX1 from specimens representative of SSU-based MOTU). Museum collection curators will need to assess the benefits and non-benefits of allowing precious specimens to be sampled for DNA, and put in place measures to prevent cross-contamination (Nadler 1999) and investigate changing storage conditions to improve DNA preservation. The databasing issue may be contentious, as attempts to coordinate taxonomic effort on this scale have previously foundered on issues of devolution versus centralization of authority, and funding (Seberg *et al.* 2003). With a sequence-based system, addition of an annexe to EMBL/GenBank that stores barcode data (and perhaps metadata as well: trace files, digital vouchers, collection data) is simple to conceive, possibly difficult but not impossible to implement, and pressing in its urgency. Freestanding efforts may also be viable (Hebert *et al.* 2003*a*,*b*). A parallel effort will have to take place to make the more traditional taxonomic literature accessible to all (Bisby *et al.* 2002; Godfray 2002*a*; Lee 2002; Oren & Stackebrandt 2002) and there will have to be coordinated effort in the molecular taxonomic community to fully investigate the within-taxon variation of barcode targets, define the between-taxon discrimination ability of targets and develop new tools for analysis of burgeoning datasets. The major contribution of a molecular taxonomy will be in throughput: the ability of a few dedicated centres to produce hundreds of thousands, and individual researchers, of molecular tags per year that can be used to diagnose species. If the biotic component of this planet is to be taxonomized, and the 'endless forms' at least listed if not understood, this is the only coherent way forward.

This paper benefited from many discussions with Robin Floyd and Eyualem Abebe, who performed the nematode barcode survey, and other members of the nematode laboratory in Edinburgh. Particular thanks go to several undergraduate project students who were guinea-pigs testing the technology of small subunit ribosomal RNA barcoding of meiofauna: Phillipa Pickles, Ronan Roche, Ingrid Iredale and Ben Elsworth. Work from the author's laboratory was funded by the Natural Environment Research Council and by the Linnaean Society of London.

REFERENCES

Adams, B. J. 1998 Species concepts and the evolutionary paradigm in modern nematology. *J. Nematol.* **30**, 1–21.

Adams, B. J. 2001 The species delimitation uncertainty principle. *J. Nematol.* **33**, 153–160.

Amaral Zettler, L. A., Gomez, F., Zettler, E., Keenan, B. G., Amils, R. & Sogin, M. L. 2002 Microbiology: eukaryotic diversity in Spain's river of fire. *Nature* **417**, 137.

Bergthorsson, U., Adams, K. L., Thomason, B. & Palmer, J. D. 2003 Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature* **424**, 197–201.

Bisby, F. A., Shimura, J., Ruggiero, M., Edwards, J. & Haeuser, C. 2002 Taxonomy, at the click of a mouse. *Nature* **418**, 367.

Blaxter, M. L. 2003 Molecular systematics: counting angels with DNA. *Nature* **421**, 122–124.

Blaxter, M. L. & Floyd, R. 2003 Molecular taxonomics for biodiversity surveys: already a reality. *Trends Ecol. Evol.* **18**, 268–269.

Blaxter, M. L. (and 11 others) 1998 A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71–75.

Blaxter, M. L., Floyd, R., Dorris, M., Eyualem, A. & De Ley, P. 2003 Utilising the new nematode phylogeny for studies of parasitism and diversity. *Nematol. Monogr. Perspect.* **2**. (In the press.)

Casamayor, E. O., Pedros-Alio, C., Muyzer, G. & Amann, R. 2002 Microheterogeneity in 16S ribosomal DNA-defined bacterial populations from a stratified planktonic environment is related to temporal changes and to ecological adaptations. *Appl. Environ. Microbiol.* **68**, 1706–1714.

Cohan, F. M. 2002 What are bacterial species? *A. Rev. Microbiol.* **56**, 457–487.

Cole, J. R. (and 10 others) 2003 The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res.* **31**, 442–443.

Dabert, J., Dabert, M. & Mironov, S. V. 2001 Phylogeny of feather mite subfamily Avenzoariinae (Acari: Analgoidea: Avenzoariidae) inferred from combined analyses of molecular and morphological data. *Mol. Phylogenet. Evol.* **20**, 124–135.

Darwin, C. 1859 *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

Daubin, V., Moran, N. A. & Ochman, H. 2003 Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832.

De Ley, P. 1999 Lost in worm space: phylogeny and morphology as road maps to nematode diversity. *Nematology* **2**, 9–16.

De Ley, P. & Bert, W. 2001 Video capture and editing as a tool for storage, distribution and illustration of morphological characters of nematodes. *J. Nematol.* **34**, 296–302.

de Meeus, T. & Renaud, F. 2002 Parasites within the new phylogeny of eukaryotes. *Trends Parasitol.* **18**, 247–251.

Diez, B., Pedros-Alio, C. & Massana, R. 2001 Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941.

Doolittle, W. F. 1999 Lateral genomics. *Trends Cell Biol.* **9**, M5–M8.

Edgcomb, V. P., Kysela, D. T., Teske, A., de Vera Gomez, A. & Sogin, M. L. 2002 Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proc. Natl Acad. Sci. USA* **99**, 7658–7662.

Eggert, L. S., Rasner, C. A. & Woodruff, D. S. 2002 The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. *Proc. R. Soc. Lond.* B **269**, 1993–2006. (DOI 10.1098/rspb.2002.2070.)

Elbadri, G. A., De Ley, P., Waeyenberge, L., Vierstraete, A., Moens, M. & Vanfleteren, J. 2002 Intraspecific variation in *Radopholus similis* isolates assessed with restriction fragment length polymorphism and DNA sequencing of the internal transcribed spacer region of the ribosomal RNA cistron. *Int. J. Parasitol.* **32**, 199–205.

Ewing, B. & Green, P. 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185.

Eyualem, A. & Blaxter, M. L. 2003 Comparison of biological, molecular and morphological methods of species Identification in a set of cultured *Panagrolaimus* isolates. *J. Nematol.* **35**, 119–128.

Felix, M.-A., Sternberg, P. W. & De Ley, P. 1996 Sinistral nematode population. *Nature* **381**, 122.

Felsenstein, J. 1978 Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.

Fitch, D. H. A., Bugaj-Gaweda, B. & Emmons, S. W. 1995 18S ribosomal gene phylogeny for some rhabditidae related to *Caenorhabditis elegans*. *Mol. Biol. Evol.* **12**, 346–358.

Floyd, R., Eyualem, A., Papert, A. & Blaxter, M. L. 2002 Molecular barcodes for soil nematode identification. *Mol. Ecol.* **11**, 839–850.

Floyd, R., Abebe, E. & Blaxter, M. L. 2004 Unveiling nematode diversity with a DNA barcode. (In preparation.)

Fuhrman, J. A., McCallum, K. & Davis, A. A. 1992 Novel major archaebacterial group from marine plankton. *Nature* **356**, 148–149.

Funch, P. & Kristensen, R. M. 1995 Cycliophora is a new phylum with affinities to Entoprocta and Ectoprocta. *Nature* **378**, 711–714.

Furlong, M. A., Singleton, D. R., Coleman, D. C. & Whitman, W. B. 2002 Molecular and culture-based analyses of prokaryotic communities from an agricultural soil and the burrows and casts of the earthworm *Lumbricus rubellus*. *Appl. Environ. Microbiol.* **68**, 1265–1279.

Gaston, K. J. & O'Neill, M. A. 2004 Automated species identification: why not? *Phil. Trans. R. Soc. Lond.* B **359**, 655–667. (DOI 10.1098/rstb.2003.1442.)

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. 1990 Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63.

Godfray, H. C. 2002*a* Challenges for taxonomy. *Nature* **417**, 17–19.

Godfray, H. C. 2002*b* Towards taxonomy's 'glorious revolution'. *Nature* **420**, 461.

Grubb, P., Groves, C. P., Dudley, J. P. & Shoshani, J. 2000 Living African elephants belong to two species: *Loxodonta africana* (Blumenbach, 1797) and *Loxodonta cyclotis* (Matschie, 1900). *Elephant* **2**, 1–4.

Hebert, P. D. N., Ratnasingham, S. & deWaard, J. R. 2003*a* Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond.* B **270**(Suppl.), S96–S99. (DOI 10.1098/rsbl.2003.0025.)

Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. 2003*b* Biological identifications through DNA barcodes. *Proc. R. Soc. Lond.* B **270**, 313–321. (DOI 10.1098/rspb.2002.2218.)

Hentschel, U., Hopke, J., Horn, M., Friedrich, A. B., Wagner, M., Hacker, J. & Moore, B. S. 2002 Molecular evidence for a uniform microbial community in sponges from different oceans. *Appl. Environ. Microbiol.* **68**, 4431–4440.

Herniou, E. A., Pearce, A. C. & Littlewood, D. T. J. 1998 Vintage helminths yield valuable molecules. *Parasitol. Today* **14**, 289–292.

Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A. & Paabo, S. 2001*a* DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799.

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Paabo, S. 2001*b* Ancient DNA. *Nature Rev. Genet.* **2**, 353–359.

Huelsenbeck, J. P. 1997 Is the Felsenstein zone a fly trap? *Syst. Biol.* **46**, 69–74.

Hugenholtz, P., Goebel, B. M. & Pace, N. R. 1998 Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774.

Karp, A. (and 11 others) 1997 Molecular technologies for biodiversity evaluation: opportunities and challenges. *Nature Biotechnol.* **15**, 625–628.

Klass, K. D., Zompro, O., Kristensen, N. P. & Adis, J. 2002 Mantophasmatodea: a new insect order with extant members in the Afrotropics. *Science* **296**, 1456–1459.

Kristensen, R. M. 1983 Loricifera, a new phylum with aschelminthes characters from the meiobenthos. *Z. zool. Syst. Evolutionsforsch.* **21**, 163–180.

Lambshead, J. 1993 Recent developments in marine benthic biodiversity research. *Oceanis* **19**, 5–24.

Lambshead, P. J. D. & Boucher, G. 2003 Marine nematode deep-sea biodiversity—hyperdiverse or hype? *J. Biogeogr.* **30**, 475–485.

Lawton, J. H. (and 12 others) 1998 Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forests. *Nature* **391**, 72–75.

Lee, M. S. 2002 Online database could end taxonomic anarchy. *Nature* **417**, 787–788.

Leser, T. D., Amenuvor, J. Z., Jensen, T. K., Lindecrona, R. H., Boye, M. & Moller, K. 2002 Culture-independent analysis of gut bacteria: the pig gastrointestinal tract microbiota revisited. *Appl. Environ. Microbiol.* **68**, 673–690.

Lipscomb, S., Platnick, N. & Wheeler, Q. 2003 The intellectual content of taxonomy: a comment on DNA taxonomy. *Trends Ecol. Evol.* **18**, 65–66.

Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C. & Moreira, D. 2001 Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607.

McDaniel, S. F. & Shaw, A. J. 2003 Phylogeographic structure and cryptic speciation in the trans-Antarctic moss *Pyrrhobryum mnioides*. *Evolution* **57**, 205–215.

Massana, R., Guillou, L., Diez, B. & Pedros-Alio, C. 2002 Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. *Appl. Environ. Microbiol.* **68**, 4554–4558.

May, R. M. 1988 How many species are there on Earth? *Science* **241**, 1441–1449.

Mercereau-Puijalon, O., Barale, J. C. & Bischoff, E. 2002 Three multigene families in *Plasmodium* parasites: facts and questions. *Int. J. Parasitol.* **32**, 1323–1344.

Moon-van der Staay, S. Y., De Wachter, R. & Vaulot, D. 2001 Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610.

Moreira, D. & Lopez-Garcia, P. 2002 The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* **10**, 31–38.

Nadler, S. A. 1999 Nucleotide sequences from vintage helminths: fine wine or vinegar? *Parasitol. Today* **15**, 122.

Nesbo, C. L., Boucher, Y. & Doolittle, W. F. 2001 Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J. Mol. Evol.* **53**, 340–350.

Nishimoto, Y., Ohnishi, O. & Hasegawa, M. 2003 Topological incongruence between nuclear and chloroplast DNA trees suggesting hybridization in the urophyllum group of the genus *Fagopyrum* (Polygonaceae). *Genes Genet. Syst.* **78**, 139–153.

Oren, A. 2004 Prokaryote diversity and taxonomy: current status and future challenges. *Phil. Trans. R. Soc. Lond.* B **359**, 623–638. (DOI 10.1098/rstb.2003.1458.)

Oren, A. & Stackebrandt, E. 2002 Prokaryote taxonomy online: challenges ahead. *Nature* **419**, 15.

Pace, N. R. 1997 A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740.

Parkinson, J., Guiliano, D. & Blaxter, M. L. 2002 Making sense of EST sequences by CLOBBing them. *BMC Bioinform.* **3**, 31.

Roca, A. L., Georgiadis, N., Pecon-Slattery, J. & O'Brien, S. J. 2001 Genetic evidence for two species of elephant in Africa. *Science* **293**, 1473–1477.

Savolainen, V., Chase, M. W., Hoot, S. B., Morton, C. M., Soltis, D. E., Bayer, C., Fay, M. F., de Bruijn, A. Y., Sullivan, S. & Qiu, Y. L. 2000 Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.* **49**, 306–362.

Seberg, O., Humphries, C. J., Knapp, S., Stevenson, D. W., Petersen, G., Scharff, N. & Andersen, N. M. 2003 Shortcuts in systematics? A commentary on DNA-based taxonomy *Trends Ecol. Evol.* **18**, 63–65.

Smith, M. W., Feng, D. F. & Doolittle, R. F. 1992 Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem. Sci.* **17**, 489–493.

Sommer, R. J., Carta, L. K., Kim, S.-Y. & Sternberg, P. W. 1996 Morphological, genetic and molecular description of *Pristionchus pacificus* sp. n. (Nematoda: Neodiplogastridae). *Fundam. Appl. Nematol.* **19**, 511–521.

Stackebrandt, E. & Goebel, B. M. 1994 Taxonomic note: a place for DNA-DNA reassociation and 16S sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol.* **44**, 846–849.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. 1996 Phylogenetic inference. In *Molecular systematics* (ed. D. M. Hillis, C. Moritz & B. K. Mable), pp. 407–514. Sunderland, MA: Sinauer Associates.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2002 DNA points the way ahead in taxonomy. *Nature* **418**, 479.

Tautz, D., Arctander, P., Minelli, A., Thomas, R. H. & Vogler, A. P. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74.

Vandenkoornhuyse, P., Husband, R., Daniell, T. J., Watson, I. J., Duck, J. M., Fitter, A. H. & Young, J. P. 2002 Arbuscular mycorrhizal community composition associated with two plant species in a grassland ecosystem. *Mol. Ecol.* **11**, 1555–1564.

Van Tienderen, P. H., de Haan, A. A., van der Linden, C. G. & Vosman, B. 2002 Biodiversity assessment using markers for ecologically important traits. *Trends Ecol. Evol.* **17**, 577–582.

Verheyen, E., Salzburger, W., Snoeks, J. & Meyer, A. 2003 Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. *Science* **300**, 325–329.

Wayne, L. G. (and 11 others) 1987 Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* **37**, 463–464.

Woese, C. R. 1996 Whither microbiology? Phylogenetic trees. *Curr. Biol.* **6**, 1060–1063.

Wuyts, J., De Rijk, P., Van de Peer, Y., Winkelmans, T. & De Wachter, R. 2001 The European large subunit ribosomal RNA database. *Nucleic Acids Res.* **29**, 175–177.

Wuyts, J., Van de Peer, Y., Winkelmans, T. & De Wachter, R. 2002 The European database on small subunit ribosomal RNA. *Nucleic Acids Res.* **30**, 183–185.

Zhou, J., Davey, M. E., Figueras, J. B., Rivkina, E., Gilichinsky, D. & Tiedje, J. M. 1997 Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. *Microbiology* **143**, 3913–3919.

GLOSSARY

MOTU: molecular operational taxonomic unit(s)
OTU: operational taxonomic unit(s)